PREDICTING CHEMICAL ODORANT TOXICITY FROM OLFACTORY BULB SCANS IN MICE

Aarya Jha
Student# 1008347406
aarya.jha@mail.utoronto.ca

Namira Kamal
Student# 1006208382
namira.kamal@mail.utoronto.ca

Ian Wu
Student# 1009063816
ii.wu@mail.utoronto.ca

Xiaofu (Fiona) Sun
Student# 1007966505
fio.sun@mail.utoronto.ca

ABSTRACT

In recent years, there has been increasing concern regarding the toxicity of chemical odorants and their potential impact on health. This project's goal is to gain insights into sensory perception and its correlation with chemical toxicity through the study of the olfactory system in mice. We propose implementing a neural network model that analyzes brain scans (748 images) of the olfactory bulb in mice. This model will predict the toxicity of chemical odorants by extracting patterns and features from brain scans. The dataset will be utilized for training, testing, and validation purposes, with 70/15/15 split. The primary model comprises Convolutional Neural Networks (CNNs) and Artificial Neural Networks (ANNs). The CNN component will extract pertinent features from the brain scans, while the ANN component will classify the toxicity levels of the chemical odorants. Through rigorous training, the model will accurately predict toxicity levels (binary classification from 0 to 1) based on the input brain scans. After training and hyperparameters search, our model produces a test accuracy of 64.29% which is higher than 62.5% produced by our selected baseline model, Random Forests. The results obtained from the neural network model will deepen our understanding of the olfactory system and contribute to advancements in the field of toxicology.

—-Total Pages: 9

1 INTRODUCTION

Capturing human capabilities within neural networks has been a feat many machine learning scientists have achieved and continue to strive to capture. Primarily this has been through creating new neural networks that mimic the way humans learn and this has provided fruitful results. Recently, there has been a push to capture human sensing capabilities in neural networks as well. Many researchers have further found that by structuring a neural network to smell like a human does, by identifying a chemical compound, the network itself begins to mimic the neuron mapping of the olfactory cortex (Burton et al., 2022).

Provided this context the team wondered what would be the possible outcomes if a neural network was trained with the input of scans of the olfactory cortex to predict what smells the brain was detecting.

Our senses, such as smell, not only inform us about our environment but also interact with our memory. This connection means that the context in which we smell something can influence how we perceive it. For instance, someone with a food allergy might perceive the smell of their allergen as toxic, whereas someone without the allergy may not. We aim to explore whether there's a relationship between how our brain identifies toxicity and different smells.

Our team believes that if the model is successful, it could be implemented in the future to better understand various psychological disorders. For instance, conditions like PTSD have begun to focus treatment developments on targeting specific portions of the brain through Transcranial Magnetic Stimulation (TMS). Neuroscientists analyze where areas of the brain are more active during the onset of a PTSD symptom and help relieve it by stimulating the area (Mahoney et al., 2020). Our model's application in this field could aid neuroscientists in uncovering connections between stimuli and their perception in the olfactory cortex. This insight might reveal links between specific smells and conditions like PTSD symptoms. Identifying and treating affected areas in the olfactory cortex could enhance our understanding of such disorders. Additionally, our model holds the potential for addressing other health conditions that require information from the olfactory cortex.

Although the team intends on looking at the future use of this model on the human brain, the scans in this project only look at the olfactory bulb, a subsection of the nervous system. Provided by Burton et al. (2022), the data is brain scans of four mice who were subjected to 185 stimuli. Using these brain scans and identifying what is toxic the team intends on creating a neural network that through input

2 ILLUSTRATION

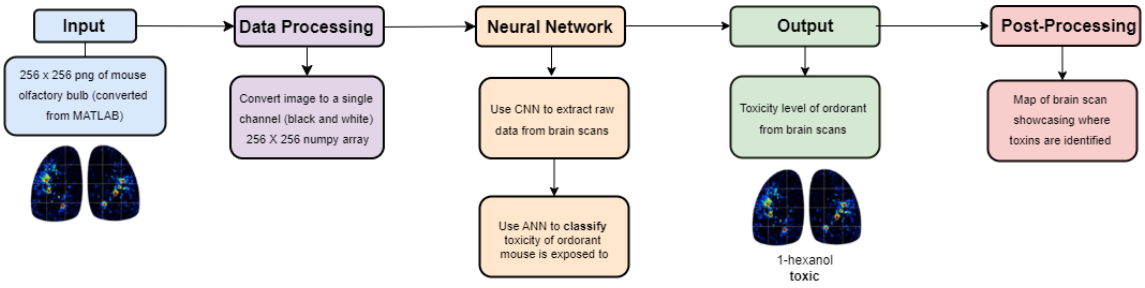


Figure 1: Illustration depicting the process of using our model to identify odorant toxicity from olfactory bulb scans of mice.

3 BACKGROUND AND RELATED WORK

We have identified five key related works that explore the use of machine learning in artificial olfaction, as well as olfaction in mice. Our project aims to combine the two ideas by using machine learning to better understand and predict olfactory behavior in mice.

3.1 Mapping odorant sensitivities reveals a sparse but structured representation of olfactory chemical space by sensory input to the mouse olfactory bulb (Burton et al., 2022)

The Atlas of Odorant Response Maps provides a comprehensive collection of response maps which are neural representations of the activity patterns observed in the olfactory system in response to 185 different odorants. These maps offer insights into the spatial organization and coding principles of olfactory information processing.

By studying the response maps, we can gain a better understanding of how odorant information is encoded and processed by the olfactory system at the neural level. This knowledge can contribute to unravelling the mechanisms underlying olfactory perception and how the brain distinguishes and recognizes different smells.

3.2 Evolving the olfactory system with machine learning (Wang et al., 2021)

This paper explores the use of machine learning techniques to evolve the olfactory system. The authors investigate the process of training a neural network to recognize and discriminate odours by leveraging genetic algorithms. By evolving the network architecture and connection strengths, they demonstrate the improvement of odour discrimination capabilities. This research presents an innovative approach to enhancing olfactory systems using machine learning and provides insights that are relevant to our project.

3.3 MoleculeNet: a benchmark for molecular machine learning (Wu et al., 2018)

This paper introduces a comprehensive benchmark dataset and evaluation framework for molecular machine-learning tasks. It provides a standardized platform to assess the performance of different machine learning models on various chemical prediction tasks, including the prediction of molecular properties, activities, and interactions.

3.4 "Electronic Nose Systems" - Alpha MOS (Hudon et al., 2000)

Alpha MOS, a leading company in the field of electronic noses, has developed commercial electronic nose systems. These devices utilize arrays of chemical sensors to capture odour profiles and employ pattern recognition algorithms to identify and classify smells. Their products are widely used in various industries, including food, beverage, and environmental monitoring, showcasing the practical application of olfactory technology.

3.5 Pathogens, odors, and disgust in rodents (Kavaliers et al., 2020)

This article discusses the role of odor-mediated pathogen disgust in rodents. It's related to our project as it explores the behaviour of mice in relation to odour detection, highlights the effects of parasite threats on animal behaviour, and focuses on the use of odours as cues to gauge the infection status of conspecifics and assess the associated threat. Additionally, the article investigates the neurobiological mechanisms underlying the pathogen disgust and the involvement of oxytocin, a nonapeptide, and steroidal hormones in the expression of this response. Furthermore, it sheds light on how these factors regulate avoidance behaviours and trade-offs in rodents. By elucidating the relationship between brain activity and behaviour in mice, this research provides valuable insights for our project.

4 DATA PROCESSING

The primary dataset this model is based on is from Burton et al. (2022) which contains 740 image scans collected from four mice. Images were taken of the mice's olfactory bulb when exposed to 185 different chemical odorants and change from baseline fluorescence was recorded. The resulting image is the average of three trials and is the input to the model. The images of the scans are not publicly available, however, one of the authors kindly provided us with the raw data used for Burton et al. (2022). Each experiment's[1] data was provided to us as a MATLAB structure with a 1x187 array, consisting of 187 256x256 single-channel scans per mouse. Using MATLAB, we export images as png files and labelled with the odorant name. The label will allow us to add toxicity information. Before exporting, the images were also rotated 90 degrees counter-clockwise to match the orientation of the scans in the original paper. In total, 740 individual images (185 for each of the four mice) will be labelled, rotated, and exported.

Since we wanted to classify odorants as toxic or non-toxic, we would need an additional dataset to determine the classification of each odorant. We accessed the U.S. Food and Drug Administration's Generally Recognized as Safe (GRAS) dataset through the Pyrfume python library (Castro et al., 2022). For each odorant, we replaced the label with a 0 or 1, based on if the odorant was present in the GRAS dataset. Additionally, the control test (no odorant exposed) was labelled "empty" and assigned a label of non-toxic (0). The results of this relabelling were stored in a new list called the modified dataset. 45% of the scans are of toxic molecules and 55% are non-toxic. Since the split was close to 50/50 we used elected not to use a confusion matrix and just used correct predictions/total predictions for accuracy.

This repository was cloned into our Colab runtime to bring the image files into our workspace. Using torch ImageFolder, we extracted the images and labelled them according to their chemical odorant. The data was first transformed into a tensor using Transform. Compose, Transform.Resize(256, 256) and Transforms.ToTensor, to ensure that each image is now a tensor of size 256x256. Since the original image contained three identical channels (same RGB values, due to it being black and white) we spliced the tensor to only include one of the channels.

We split the data into 70% training, 15% validation, and 15% test. Each bucket will contain a proportional amount of toxic and non-toxic labelled molecules. For example, 87/185 molecules are on the Generally Recognized As Safe list, so each bucket will roughly have the same ratio of toxic: non-toxic scans. To show that your data processing was complete, a sample Dataloader was created and a training image was shown through matplotlib.imshow(). Note that we did not normalize the image pixels to [0, 1] as they were already normalized when we received them in MATLAB.

At this point, all data has been processed and the arrays, the ground truth labels are set, and the Dataloaders are made, it is ready to be used in training the neural network. This process can be seen in more detail through the Colab file in section13.

5 ARCHITECTURE

The model consists of a Convolutional Neural Network (CNN) whose output is then fed into an Artificial Neural Network (ANN). This configuration was chosen since our project was focused on classifying images of mouse brain scans as toxic or non-toxic. CNN's are notable for being better at handling images as the convolutional layers are able to reduce the dimensionality of the images while containing important information, allowing our model to extract important information about the features of the brain scans. These features are then flattened and passed into an ANN. The ANN then helps to classify whether the images showcase a brain that has been exposed to a toxic or non-toxic chemical.

[1]Note that the raw data has information for 8 experiments since the paper examines the left and right bulbs separately. However, they are captured in the same scan, so for this model, only 4/8 experiments will be exported

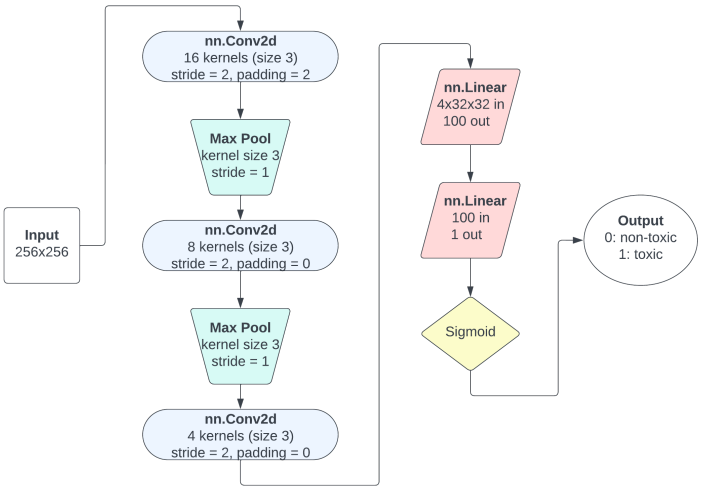


Figure 2: The model architecture expressed as a flowchart, CNN and ANN sections have been separated into columns.

In class PrimaryModel the first segment consists of the CNN which consists of a convolutional layer that takes in the input of the single channel image and produces an output of size 16 with a kernel of size 3, stride of size 2 and the standard padding of size 2. The information is then passed into a ReLU activation function which was chosen as it was compatible with the outputs of the convolutional layer and would help decrease non-linearity. The output of this is then passed into the MaxPool2d function to help extract important features and remove invariances. This is passed into another convolutional layer with an output size of 8, with the same kernel and padding but the stride size is now reduced to 1. The output of this layer passes through the ReLU function and MaxPool2d function as well, and is inputted into the final convolutional layer with the same kernel, stride and padding size as the first producing an output of size 4 x 32 x 32. In the CNN there is a total of 252 trainable parameters. These features are flattened and passed into the second segment of PrimaryModel which is the ANN. The ANN consists of 2 linear layers, the first taking an input of 4 x 32 x 32 and producing an output of size 100. This is then passed into the ReLU activation and passed into the final Linear layer. The team did not add an output sigmoid activation function as this was done through the training module through BCEWithLogitsLoss. The ANN had a total of 409700 trainable parameters. Overall the Primary Model had 409952 trainable parameters.

Lastly, the optimal hyperparameters were selected using a hyperparameter search to ensure the model perform the best accuracy. It is determined using a batch size of 256, a learning rate of 0.0005, momentum of 0.9 will produce the best accuracy of around 0.703 at epoch 142.

6 Baseline Model

In an attempt to create an accurate and robust predictive model, our team experimented with three different baseline models. We selected the Random Forest model as our reference. Each model was chosen based on its unique strengths and applicability to our dataset, with the aim to explore a wide range of machine learning techniques. The baseline models, Support Vector Machines (SVM), Naive Bayes, and Random Forest were trained and tested on our data. Each model's performance was evaluated using validation and

testing datasets, allowing us to understand their predictive prowess and identify any areas of overfitting or underperformance, allowing for a strong comparison with the primary model. Considering both the performance and the characteristics of the models, the Random Forest model is selected as it achieved the highest accuracy on the test set, is robust to outliers, can handle non-linear data, and provides insight into feature importance. Random Forest is a machine learning algorithm that combines multiple decision trees to generate robust predictions. The key strength of this model is its ability to manage high-dimensional datasets efficiently, ensuring higher accuracy and providing important insights into feature significance (Pedregosa et al., 2011).In our code, the Random Forest model is instantiated with 300 estimators (or trees), each contributing to the final prediction. The same data flattening process was applied to the training, validation, and testing datasets, as in the SVM model. After the model was trained using the fit method, it was evaluated on the validation set, producing an accuracy score of 60.3%. The model was further evaluated on the test set, achieving an accuracy of 62.5%.

7 Quantitative Results

Accuracy was chosen as our primary metric for evaluating the performance of our model as Accuracy is particularly suitable for balanced classification problems, which is characteristic of our dataset. Our primary model shows promising results with a validation accuracy of 71.1% and a test accuracy of 68.5%. These results are consistent with multiple runs of our model.

The primary model performs significantly better than the baseline which had a test accuracy of around 62.5% and a validation accuracy of 60.3%, we did notice an offset in our training validation start due to a random split between toxic and non-toxic data, which is something we plan to investigate further.

As part of our quantitative measurements, we also observed the model's learning process closely. As shown in Figure A we implemented early stopping at 150 epochs to prevent overfitting, but the model still had some difficulty achieving convergence, which is something we aim to improve.

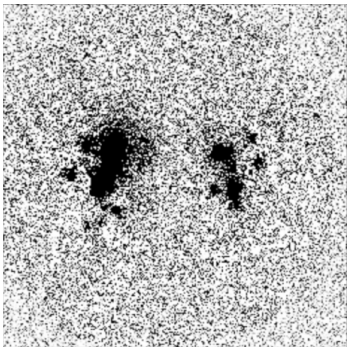
8 Qualitative Results

We present two examples of correctly labelled images by the model in Figure 3 and two examples of incorrectly labelled images in Figure 4. An example of each aspect of the confusion matrix is shown to try to better understand the strengths and shortcomings of the model. However, note the small sample size. Also note, that though these are qualitative results, the abstract nature of these scans according to Burton et al. (2022), makes it difficult to determine the patterns resulting in success or failure.

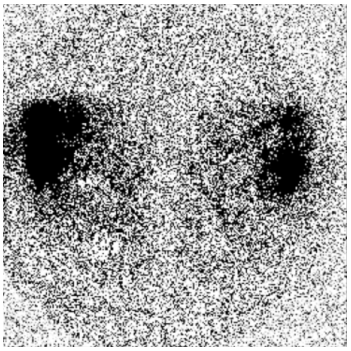
9 Evaluating Model on New Data

The final model was tested on a test set containing 112 unique, never before seen, images that were randomly split from the original citet/burton dataset. The model performed at an accuracy of 64.29% on the test data. This outperforms the peak baseline model developed using Random Forests which scored 62.5%. A true positive rate of 32.56% and a true negative rate of 84.06% was achieved. This model was better at negative labeling but struggled in positive labeling.

An attempt was made to test the model on data outside of the original dataset. However, the novelty of the research and specific imaging process made it difficult to find comparable data to test the model on. We were able to find another study that used a very similar method and produced similar scans, but only 5 odorants were used producing 5 scans. Due to the extremely small data, we concluded it would be statistically insignificant to test on a separate dataset.

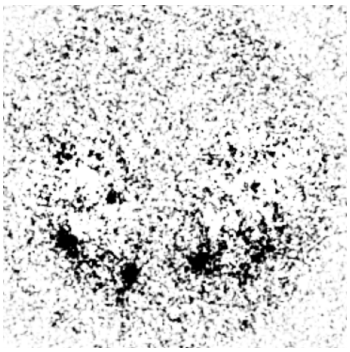


(a) A toxic odorant scan

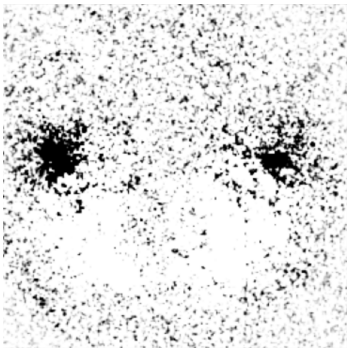


(b) A non-toxic odorant scan

Figure 3: Correctly labelled scans by the model.



(a) A toxic odorant scan



(b) A non-toxic odorant scan

Figure 4: Incorrectly labelled scans by the model.

10 DISCUSSION

The model's performance is adequate achieving a validation accuracy of 71.1%, and a testing accuracy of 64.29%. Ideally, these numbers would be much higher, however, these results were better than our baseline model.

Our model seemed to have difficulty converging. The results depict that the data would converge rather rapidly. This would result in tendencies of underfitting or overfitting during the training runs.

When looking at section 8 it appears that the model performed best on higher-quality with lower contrast between light and dark spots seen in Figure 3. The two incorrectly labelled images appear to have exceeded the range of the scanning device resulting in dark darks and bright white spots.

Seeing the true positive rate and true negative rate, the presence of underfitting is obvious. The model tends to prefer to simply guess one label over the other more often, suggesting it has not been properly trained. This became a very difficult thing to balance between underfit and overfit.

A significant issue is the lack of available training data. In an ideal training scenario, we could train the model for several epochs, but a small dataset requires early stopping to prevent overfitting. Thus we found

our model was incapable of forming deeper connections and stronger inferences. Even though this problem is not severely complicated, the limited resources severely hampers the performance of the model.

A minor inconvenience was our images being a single channel. To further understand the flaws in our model the kernels were extracted from each of the convolutional layers. The image is first inverted and then in the following kennels the image goes through a series of grey-scaling, varying the brightness and exposure, ultimately ending with a gray image where the areas of interest are smoother.

In these images, the flaws of having single-channel images are apparent. In the final kernel produced, the pixels are all shades of gray varying slightly in value, although the area of interest has more homogenous pixels, there does seem to be more noise within the image overall. This could potentially have been avoided if the images were multi-channel.

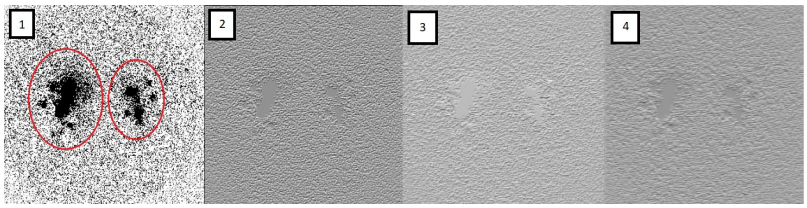


Figure 5: Illustration depicts the kernels of the brain scan image after each convolutional layer

Lastly, the merging of two datasets (scans and toxicity labels) may not be perfect, which could be affecting the accuracy of results and cause them to be lower. The model is trained on the assumption that the ground truth labels are 100% accurate. This assumption is likely false, however, as the toxicity evaluation was done through a study that can have mistakes.

11 Ethical Considerations

While the proposed neural network holds promising potential for advancing our understanding of the olfactory system in mice and enhancing smell recognition models, its application could raise ethical concerns. One possible use of the system that could give rise to ethical issues is its utilization in the development of new toxic substances or odour-based weapons. By accurately identifying scents and assessing their toxicity, there is a risk that this technology could be exploited to create harmful compounds specifically designed to evade traditional detection methods. Such misuse could have serious implications for public safety and security, as well as raise questions about the responsible use of advanced neural network technologies.

Despite its potential, the model does have limitations. The neural network's performance heavily relies on the quality and diversity of the training data. If the dataset used to train the network is limited in terms of the types of odours and toxicity levels, the model might struggle to generalize effectively to new and diverse scents or toxicity profiles. Additionally, translating findings from mice to humans is a complex endeavour due to inherent biological differences between species. This raises concerns about the applicability of the insights gained from mouse olfactory scans to human sensory perception. Moreover, our team has taken into account the ethical considerations regarding the use of animal subjects and the potential harm they may experience during brain scan data collection. We ensured that the data was collected following all protocols and that animals were treated according to the National Institutes of Health Guide for the Care and Use of Laboratory Animals. However, using animal data as a necessary part of our model should be considered as a limitation as when others try to reproduce our model in another set of experiment data, they might not be as ethical as us. Addressing these limitations and ethical concerns will be crucial in ensuring the responsible and effective deployment of our neural network technology.

12 PROJECT DIFFICULTY

The team initially had the intention to find a problem that was compatible with our current knowledge of deep learning at the time. The problem would need solution whose model could work on CNN and ANN model architectures. Our problem worked perfectly as could take brain scans of the mice exposed to specific odorants, passing them into a CNN to extract useful features that could then be passed into an ANN to help with the classification of the brain scan to its respective odorant and toxicity level. The problem had an added challenge to it as the CNN was connected to the ANN and the model had to classify the image into two different label categories.

The difficulty of our problem increased as we worked on it. We were provided with MATLAB Arrays that were then converted into single channel 256 by 256 Numpy Arrays. This was fed into CNN. In attempting to extract meaningful features of the kernels by varying the intensity of the pixels, final kernel produced a more nosier image as in the discussion section of the report. The CNN would have likely performed better in the image classification processes if the images were multi-channel.

To improve our testing and validation score and combat the issues we decided to implement different versions of our original model. We tried the following and obtained validation and testing accuracy's lower than the scores achieved by our current model: Changing Hyperparameters/Layers of Current Model, ADAM as an Optimizer Instead of SGD, An Autoencoder with Convolutional Layers Fed into an ANN, Transfer Learning

Each of these changes presented a testing and validation score lower than our current model, which is why they were not included. To better understand why we were facing difficulty in improving our score we looked more into the research conducted by Burton on the mice. Burton pointed out, the glomerulus-odorant response matrices in our data are high-dimensional. This complexity, not yet fully understood by humans, may have impeded the model's learning. Working with a smaller dataset also posed a challenge to our model. With a larger dataset, our performance could improve drastically as we could train it for longer with larger epochs with less risk of over or under fitting.

Despite the challenges, after re-tweaking our model we were able to provide adequate results, showing that perhaps by looking at other models that could handle this level of dimensionality we could potentially obtain a better score.

13 GOOGLE COLAB LINK (MODEL CODE) AND GITHUB (DATA SOURCE)

`https://colab.research.google.com/drive/1p5JjkQUAp-F9KlMppHTEQmtohMaROP6z?usp=sharing`

`https://github.com/bunnyian/mouse-olfactory-scans`

REFERENCES

Shawn D Burton, Audrey Brown, Thomas P Eiting, Isaac A Youngstrom, Thomas C Rust, Michael Schmuker, and Matt Wachowiak. Mapping odorant sensitivities reveals a sparse but structured representation of olfactory chemical space by sensory input to the mouse olfactory bulb. *eLife*, 11:e80470, jul 2022.

Jason Castro, Travis Gould, Robert Pellegrino, Zhiwei Liang, Liyah Coleman, Famesh Patel, Derek Wallace, Tanushri Bhatnagar, Joel Mainland, and Richard Gerkin. Pyrfume: A window to the world's olfactory data. 09 2022.

Guillaume Hudon, Christophe Guy, and Jacques Hermia. Measurement of odor intensity by an electronic nose. *Journal of the Air & Waste Management Association*, 50(10):1750–1758, 2000.

Martin Kavaliers, Klaus-Peter Ossenkopp, and Elena Choleris. Pathogens, odors, and disgust in rodents. *Neurosci Biobehav Rev*, 119:281–293, Dec 2020.

J. J. Mahoney, C. A. Hanlon, P. J. Marshalek, A. R. Rezai, and L. Krinke. Transcranial magnetic stimulation, deep brain stimulation, and other forms of neuromodulation for substance use disorders: Review of modalities and implications for treatment. *Journal of the Neurological Sciences*, 418:117149, Nov 2020. doi: 10.1016/j.jns.2020.117149.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peter Y Wang, Yi Sun, Richard Axel, L F Abbott, and Guangyu Robert Yang. Evolving the olfactory system with machine learning. *Neuron*, 109(23):3879–3892, Dec 2021.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9: 513–530, 2018.